Abstract of the Disclosure

The present invention relates to a method and system for providing Web content from pull and push based services running on Web content providers to mobile users. A proxy gateway connects the mobile users to the Web content providers. A prefetching module is used at the proxy gateway to optimize performance of the pull services by reducing average access latency. The average access latency can be reduced by using at least three factors: one related to the frequency of access to the pull content; second, the update cycle of the pull content determined by the Web content providers; and third, the response delay for fetching pull content from the content provider to the proxy gateway. Pull content, such as documents, having the greatest average access latency are sorted and a predetermined number of the documents are prefetched into the cache. Push services are optimized by iteratively estimating a state of each of the mobile users to determine relevant push content to be forward to the mobile user.